Statistique élémentaire avec **R**
Partie 1 : Probabilités, Statistiques Descriptives, Estimation

Julien JACQUES

Université Lumière Lyon 2

Plan

Plan

Probabilités

Le formalisme

- $\mathcal{E}$: expérience aléatoire
- Ω : ensemble des résultats possibles de $\mathcal{E}$ (*univers*)
- $A \in \Omega$: événement

Opérations sur les événements

- union $A \cup B$: A ou B (ou les 2) est réalisé
- intersection $A \cap B$: A et B sont réalisés en même temps
  si $A \cap B = \emptyset$, A et B sont incompatibles
- complémentaire $\bar{A}$: A n'est pas réalisé

Propriété : $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

Probabilités

Définition d'une probabilité

Probabilité :

$$\begin{array}{rcl} P : \Omega & \rightarrow & [0,1] \\ A & \rightarrow & P(A) \end{array}$$

où $P(A) \simeq$ évaluation des chances de réalisation de A lors d'une expérience aléatoire.

- $0 \leq P(A) \leq 1$ pour tout $A \in \Omega$
- $P(\Omega) = 1$
- $A, B \in \Omega$ tels que $A \cap B = \emptyset$: $P(A \cup B) = P(A) + P(B)$
  on dit que A et B sont *incompatibles*

Probabilités

Propriétés

- $P(\emptyset) = 0$
- $P(\bar{A}) = 1 - P(A)$
- si $A \subset B$: $P(A) \leq P(B)$
- si $A \cap B \neq \emptyset$: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Quelques définitions

- probabilité *conditionnelle* que A soit réalisé sachant que B l'est

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- *indépendance* : A et B sont indépendants si $P(A|B) = P(A)$
  $\Rightarrow P(A \cap B) = P(A)P(B)$

Probabilités

Formule des probabilités totales

Soit $A_1, A_2, \ldots, A_n$ un système complets d'événements, i.e. :

- $A_i \cap A_j = \emptyset$ pour tout $i \neq j$,
- $A_1 \cup A_2 \cup \ldots \cup A_n = \Omega$.

Alors pour tout B :

$$P(B) = \sum_{i=1}^{n} P(B \cap A_i)$$

Probabilités

Formules de Bayes

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

et si $A_1, A_2, \ldots, A_n$ est un système complets d'événements

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{i=1}^{n} P(B \cap A_i)} = \frac{P(B|A_i)P(A_i)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}$$

Probabilités

Calculer une probabilité

Lorsqu'on sait dénombrer les événements :

$$P(A) = \frac{\text{nombre de cas favorables}}{\text{nombre de cas possibles}}$$

valide lorsque tous les cas sont équiprobables

Dénombrement

- $n! = 1 \times 2 \times \ldots \times n$ (nombre de permutations de n objets) : nombre de façons de ranger n objets
- C_n^k (combinaisons de k parmi n) : nombre de choix possibles de k objets parmi n

$$C_n^k = \frac{n!}{k!(n-k)!}$$

Variables aléatoires

Variable aléatoire

$$X : \Omega \to E$$

- Ω : ensemble d'événements possibles ou d'individus
- E : valeurs possibles de X

Les différents types de variables aléatoires

- quantitative
  - discrète : nombre de clients $E = \{1, 2, 3, ...\}$
  - continue : taille, poids $E = \mathbb{R}$
- qualitative
  - nominale : couleur des yeux $E = \{bleu, vert, noir, ...\}$
  - ordinale : mention au BAC $E = \{P, AB, B, TB...\}$

Plan

Variables aléatoires quantitatives discrètes

$$X : \Omega \to E \qquad \text{avec } E = \{x_1, \ldots, x_k\}$$

Distribution (loi) de X

$$X \sim \begin{pmatrix} x_1 & \ldots & x_i & \ldots & x_k \\ p_1 & \ldots & p_i & \ldots & p_k \end{pmatrix}$$

où $p_i = P(X = x_i)$ $\qquad (\sum_{i=1}^{k} p_i = 1)$

On supposera (par convention)

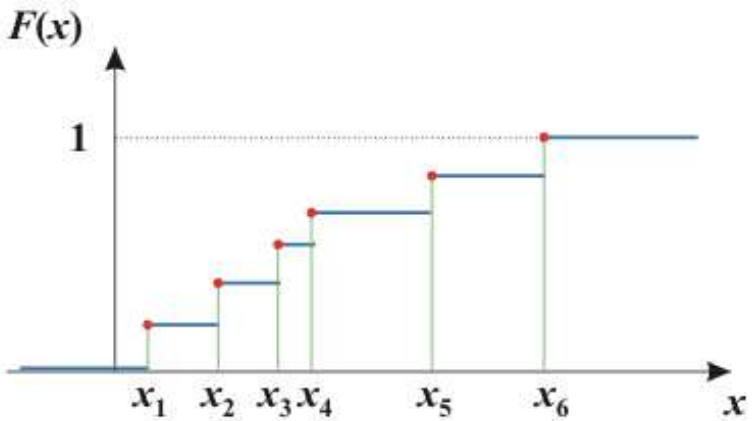
$$x_1 \leq \ldots \leq x_i \leq \ldots \leq x_k$$

.

Caractéristiques d'une variable discrète

Fonction de répartition

$$F_X(x) = P(X \leq x) = \begin{cases} 0 & \text{si } x < x_1 \\ p_1 & \text{si } x_1 \leq x < x_2 \\ p_1 + p_2 & \text{si } x_2 \leq x < x_3 \\ & \vdots \\ 1 & \text{si } x \geq x_k \end{cases}$$



Caractéristiques d'une variable discrète

Espérance

$$E[X] = \sum_{i=1}^{k} p_i x_i$$

valeur théorique moyenne à laquelle on peut s'attendre pour X

Propriétés

- on note $E[X] = \mu$
- si X et Y sont deux variables aléatoires, et a une constante : $E[aX + Y + b] = aE[X] + E[Y] + b$
- $E[XY] \neq E[X]E[Y]$ sauf si X et Y sont indépendantes

Caractéristiques d'une variable discrète

Variance

$$V(X) = E[(X - E[X])^2] = \sum_{i=1}^{k} p_i (x_i - \mu)^2 = \sum_{i=1}^{k} p_i x_i^2 - \mu^2$$

moyenne des carrés des écarts de X à sa moyenne μ

Propriétés :

- on note $V(X) = \sigma^2$
- on appelle écart-type : $\sigma = \sqrt{V(X)}$
- $V(a + X) = V(X)$ et $V(aX) = a^2 V(X)$
- $V(X + Y) \neq V(X) + V(Y)$ sauf si X et Y sont indépendantes

Exemples de lois discrètes

Soit $A \in \Omega$ tel que $P(A) = p$.

Bernoulli $\mathcal{B}(p)$

On réalise une expérience aléatoire $\mathcal{E}$ et on définit

$$X = 1 \text{ si } A \text{ est réalisé, 0 sinon}$$

$$X \sim \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}$$

- $E[X] = p$
- $V(X) = p(1-p)$

Exemples de lois discrètes

Soit $A \in \Omega$ tel que $P(A) = p$.

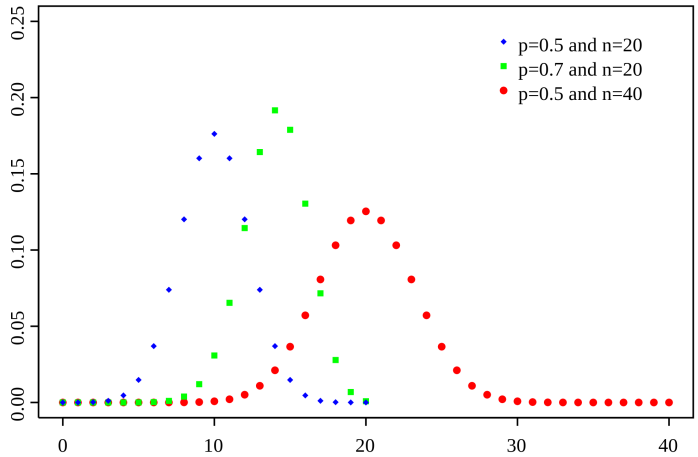
binomiale $\mathcal{B}(n, p)$

On répète n fois de façon indépendante et dans les mêmes conditions l'expérience aléatoire $\mathcal{E}$

$X =$ nombre de fois où A est réalisé.

$$X \sim \begin{pmatrix} 0 & 1 & \dots & n \\ p_0 & p_1 \dots & & p_n \end{pmatrix}$$

avec $p_j = P(X = j) = C_n^j p^j (1-p)^{n-j}$



Propriétés:

- $E[X] = np$
- $V(X) = np(1-p)$

Exemples de lois discrètes

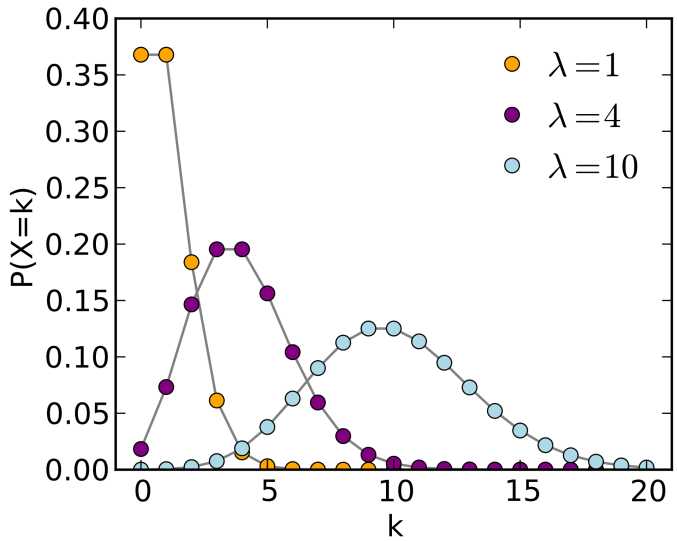
Poisson $\mathcal{P}(\lambda)$

On compte un nombre d'événements A indépendants se produisant dans un intervalle de temps fixé

X = nombre d'événements A réalisés.
$X \in \mathbb{N}$
avec $P(X = j) = e^{-\lambda} \frac{\lambda^j}{j!}$

Propriétés:

- $E[X] = \lambda$
- $V(X) = \lambda$



Exercice sur la loi binomiale

Lancer de dé

- je lance un dé non truqué. Quelle est la probabilité d'obtenir 6 ?
- je lance deux dés non truqués. Quelle est la probabilité d'obtenir au moins un 6 ?
- je lance dix dés non truqués. Quelle est la probabilité d'obtenir au moins un 6 ?

Plan

Variables aléatoires quantitatives continues

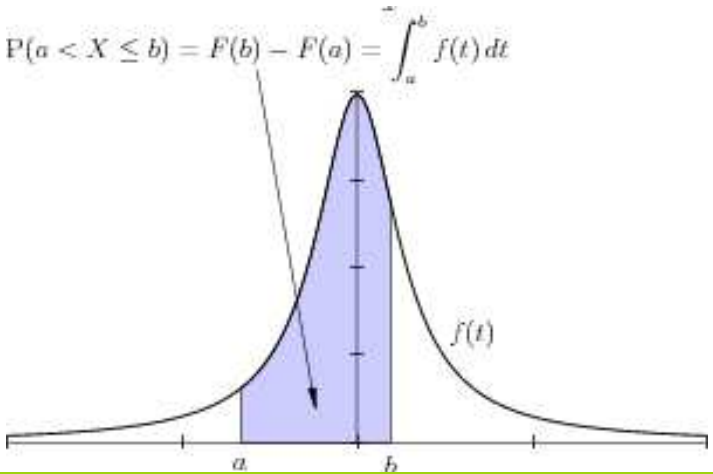
$$X : \Omega \to E \qquad \text{avec } E \subseteq \mathbb{R}$$

$P(X = x) = 0$ pour tout $x \in E$ car E non dénombrable.
On s'intéresse plutôt à $P(X \in [a, b])$.

Distribution et fonction de densité

X est caractérisée par sa fonction de densité $f_X : \mathbb{R} \to \mathbb{R}^+$

$$P(X \in [a, b]) = P(X \in]a, b]) = \int_a^b f_X(x)dx$$



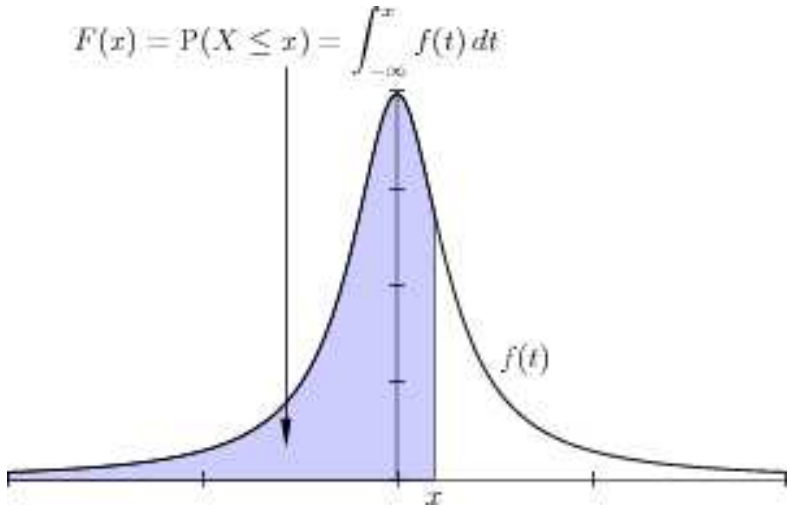
Variables aléatoires quantitatives continues

$$X : \Omega \to E \qquad \text{avec } E \subseteq \mathbb{R}$$

Fonction de répartition

X peut également être caractérisée par sa fonction de répartition
$F_X : \mathbb{R} \to [0, 1]$

$$P(X \leq a) = P(X < a) = \int_{-\infty}^{a} f_X(x)dx$$



Caractéristiques d'une variable continue

Espérance

$$E[X] = \int_{-\infty}^{+\infty} x\, f_X(x) dx = \mu$$

valeur théorique moyenne à laquelle on peut s'attendre pour X

De plus, pour *toute* fonction h : $E[h(X)] = \int_{-\infty}^{+\infty} h(x) f_X(x) dx$

Variance

$$V(X) = \int_{-\infty}^{+\infty} (x-\mu)^2 f_X(x) dx = \int_{-\infty}^{+\infty} x^2 f_X(x) dx - \mu^2$$

moyenne des carrés des écarts de X à sa moyenne μ

Caractéristiques d'une variable continue

Quantile

On appelle quantile d'ordre α, le nombre q_α tel que :

$$P(X \leq q_\alpha) = \alpha \quad \text{et} \quad P(X \leq q_\alpha) = 1 - \alpha$$

- Pour une loi discrète, il peut exister une infinité de valeur respectant ces deux conditions.
  On prendra par convention la plus petite.
- Pour une loi continue, les deux égalités sont vérifiées.

Médiane

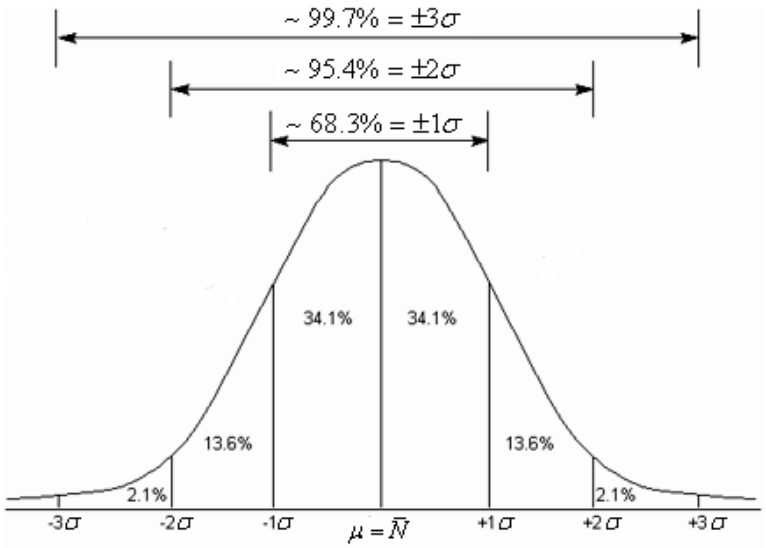
On appelle médiane le quantile d'ordre $\alpha = 50\%$.

Exemple de loi continue : la loi normale

normale $\mathcal{N}(\mu, \sigma^2)$

Densité de la loi normale (gaussienne) de paramètres μ et σ^2 :

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$



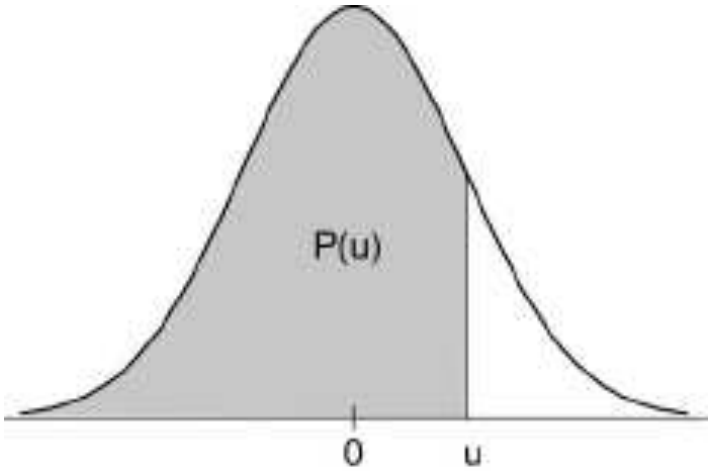
Propriétés :

- $E[X] = \mu$
- $V(X) = \sigma^2$
- symétrie :
  $P(X < \mu - a) = P(X > \mu + a)$

Exemple de loi continue : la loi normale

Cas particulier normale centrée réduite $\mathcal{N}(0,1)$

Il existe des tables permettant de lire $P(X < u)$ pour $u > 0$.
La fonction `pnorm` de R donne également $P(X < u)$ pour u donné.



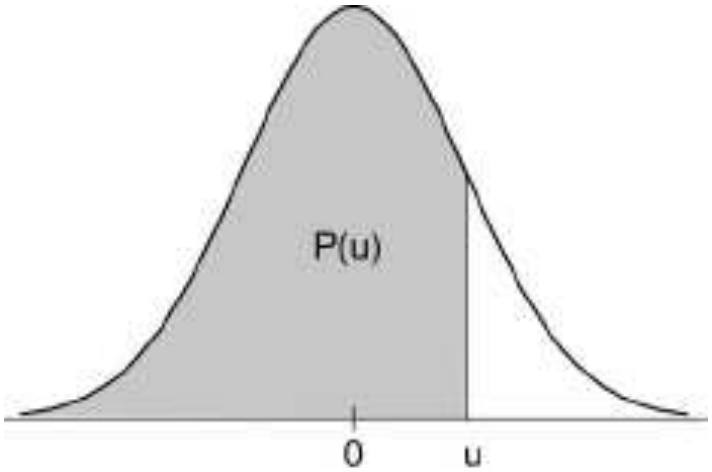
Exemples :

$$P(X < 1.6449) = 95\%$$
$$P(X < 1.96) = 97.5\%$$

Exemple de loi continue : la loi normale

Cas particulier normale centrée réduite $\mathcal{N}(0,1)$

La fonction `qnorm` donne la valeur du quantile d'ordre α, souvent notée u_α pour la loi normale centrée réduite.



Exemple de loi continue : la loi normale

Utilité de la $\mathcal{N}(0,1)$

Si $X \sim \mathcal{N}(\mu, \sigma^2)$

$$Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0,1)$$

d'où

$$P(X < a) = P\left(\frac{X - \mu}{\sigma} < \frac{a - \mu}{\sigma}\right) = P\left(Z < \frac{a - \mu}{\sigma}\right)$$

et on utilise la table.

Exercice sur la loi normale

Calcul de probabilité normale

Soit $X \sim \mathcal{N}(2, 4)$. Calculer, à l'aide des tables statistiques

- $P(X = 2)$
- $P(X > 2)$ et $P(X \geq 2)$
- $P(0 < X < 4)$ et $P(-2 < X < 6)$
- Soit $Y \sim \mathcal{N}(0, 1)$. Quel lien y-a-t'il entre $P(X > 4)$ et $P(Y > 1)$?

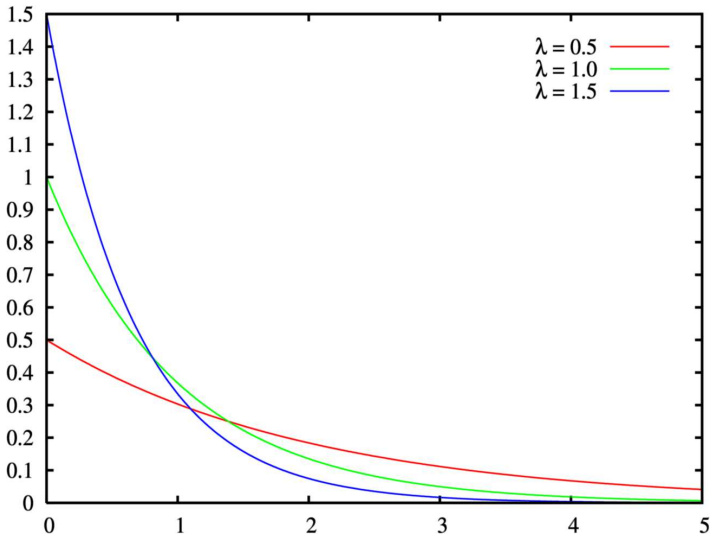
Exemple de loi continue : la loi exponentielle

exponentielle $\mathcal{E}(\lambda)$

Modélise des durées (de vie en fiabilité, pour matériel sans usure ni fatigue)

Densité de la loi exponentielle de paramètre λ :

$$f_X(x) = \lambda e^{-\lambda x} \text{ si } x > 0$$



Propriétés:

- $E[X] = 1/\lambda$
- $V(X) = 1/\lambda^2$

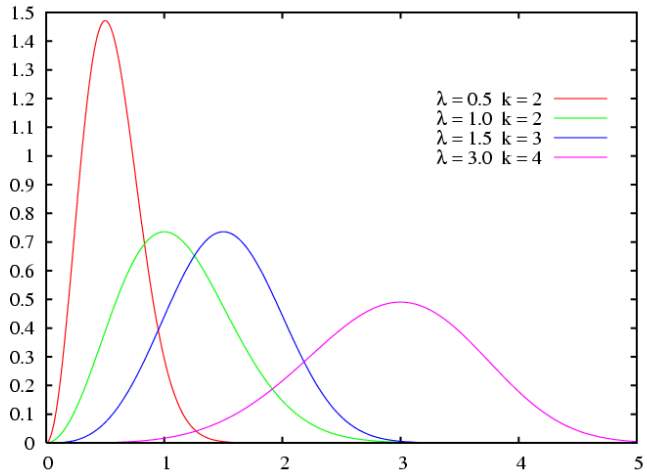
Exemple de loi continue : la loi de Weibull

Weibull $\mathcal{W}(\alpha, \beta)$

Modélise des durées de vie en fiabilité

Densité de la loi de Weibull de paramètre α, β :

$$f_X(x) = \frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1} \lambda e^{-(x/\beta)^{\alpha}} \text{ si } x > 0$$



Propriétés:

- α lié au vieillissement:

$\alpha = 1 \Rightarrow$ loi $\mathcal{E}(1/\beta)$,

$\alpha > 1 \Rightarrow$ matériel qui se fatigue

- β lié à la durée de vie médiane:

$\beta = \textit{mediane}/((\ln 2)^{1/\alpha})$

Exemple de loi continue : la loi de Gumbel

Gumbel standard $\mathcal{G}$

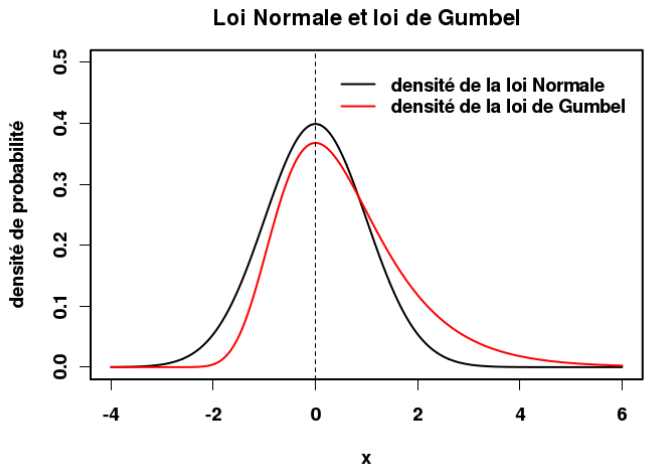
Modélise des valeurs extrêmes (crues/magnitudes maximales annuelles...)

Densité :

$$f_X(x) = e^{-x-e^{-x}} \text{ si } x > 0$$

Fonction de répartition :

$$F_X(x) = p(X < x) = e^{-e^{-x}} \text{ si } x > 0$$



Plan

Liaison entre deux variables aléatoires quantitatives

Covariance

On définit la covariance entre deux variables aléatoires X et Y par

$$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - \mu_X \mu_Y.$$

Interprétation :
Exprime la présence d'une relation du type $Y = aX + b$

- $Cov(X, Y) > 0$: les variables varient dans le même sens par rapport à leur moyenne,
- $Cov(X, Y) < 0$: les variables varient en sens inverse,
- $Cov(X, Y) = 0$: les variables ne sont pas corrélées linéairement (Attention : cela n'implique pas l'indépendance (sauf si les variables sont normales)

Remarque : $Cov(X, X) = V(X)$

Liaison entre deux variables aléatoires quantitatives

Corrélation

Comme la covariance n'est pas bornée, on définit la corrélation entre deux variables aléatoires X et Y par

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}.$$

Propriété : $\rho_{X,Y} \in [-1,1]$

Liaison entre deux variables aléatoires quantitatives

Indépendance

Deux variables aléatoires X et Y sont indépendantes si

$$P(\{X \in A\} \cap \{Y \in B\}) = P(\{X \in A\})P(\{Y \in B\}).$$

Propriétés :

- si X et Y sont à densité,
  X et Y indépendantes $\Rightarrow f_{X,Y}(x,y) = f_X(x)f_Y(y)$
- X et Y indépendantes $\Rightarrow Cov(X,Y) = 0 = \rho_{X,Y}$
  (Attention $\not\Leftarrow$)
- X et Y indépendantes $\Rightarrow$
  $f_{X+Y}(z) = f_X(x) \otimes f_Y(y) = \int_{\mathbb{R}} f_X(u)f_Y(z-u)du$

Liaison entre deux variables aléatoires quantitatives

Calcul d'inégalité aléatoire

Si X et Y sont indépendantes

$$\begin{aligned} P(X > Y) &= \int_{-\infty}^{\infty} \int_{-\infty}^{x} f_Y(y) f_X(x) dx dy \\ &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{x} f_Y(y) dy \right) f_X(x) dx \end{aligned}$$

Exemple d'application :

- X : resistance d'un matériau
- Y : force appliquée sur le matériau
- $P(X > Y)$: probabilité que le matériau résiste

Plan

Notion d'échantillon

Soit X la variable aléatoire d'intérêt.

Échantillon

On appelle échantillon toute suite $X_1, \ldots, X_n$ de variables aléatoires indépendantes et de même loi que la variable aléatoire X d'intérêt.

Exemple :
X : taille des individus de la population française
$X_1, \ldots, X_n$: taille de n individus choisis au hasard dans la population française.

On notera $x_1, \ldots, x_n$ l'observation de l'échantillon aléatoire $X_1, \ldots, X_n$.

Plan

Résumés numériques d'une variable quantitative

Tendance centrale

- Moyenne : $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$
- Médiane : valeur M qui partage l'échantillon ordonné en 2

$$M = X_{\frac{n+1}{2}} \text{ si n impair, et } M = \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2}$$

en supposant que l'échantillon $X_1, \ldots, X_n$ est classé par ordre croissant

- Mode : valeur la plus observée (*pour variables discrètes*)

Rq : $\bar{X}$ et M sont des variables aléatoires, leurs observations sur un échantillon sont notées $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ et m.

Résumés numériques d'une variable quantitative

Dispersion

- Variance (empirique) : $V^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$
- Étendue : $X_{max} - X_{min}$
- Intervalle inter-quartile : $[Q_1, Q_3]$
  où le premier quartile Q_1 est la valeur telle qu'1/4 des observations lui soient inférieures et 3/4 supérieures (et l'inverse pour le troisième quartile Q_3)

Résumés numériques d'une variable quantitative

Forme

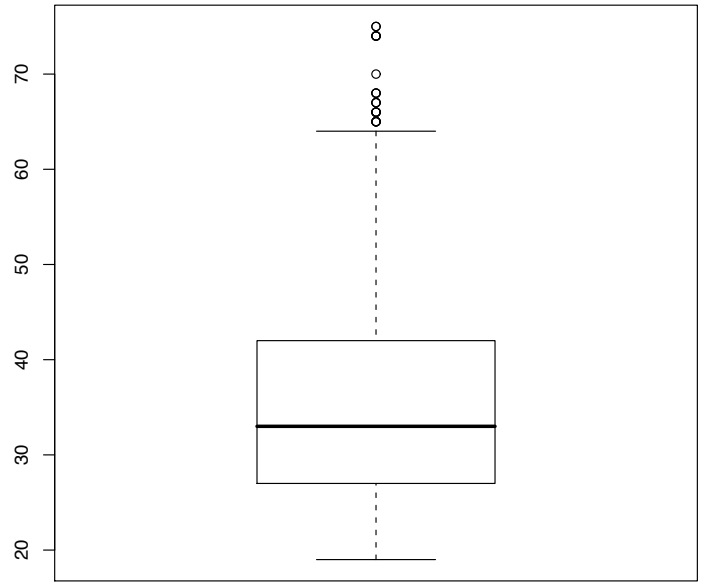
- skewness : coefficient d'asymétrie $\gamma_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^3}{(\sqrt{n/(n-1)}\,V)^3}$
- kurtosis : coefficient aplatissement $\gamma_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^4}{(n/(n-1))^2\,V^4}$

Propriétés :

- $X \sim \mathcal{N} \Rightarrow \gamma_1 = 0$ et $\gamma_2 = 3$ (0 pour Excel)
- $\gamma_1 > 0$: distribution décalée vers la gauche
- $\gamma_2 > 3$: distribution plus aplatie qu'une gaussienne

Représentation graphique d'une variable quantitative

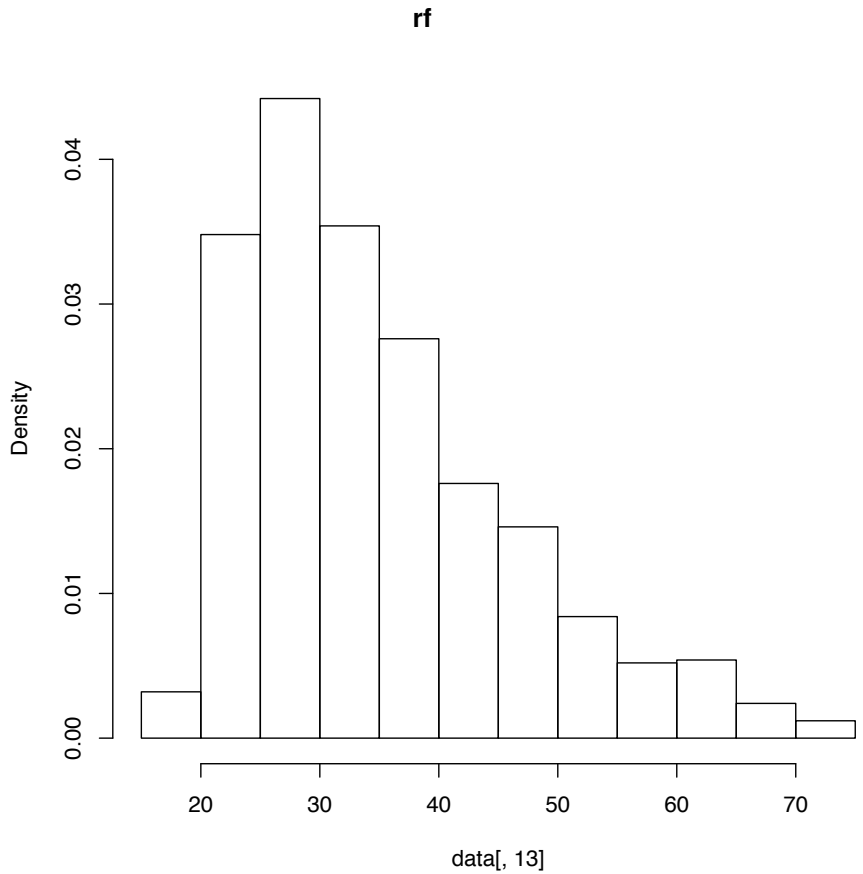
Boîte à moustaches



- trait centré au milieu de la boîte : médiane
- la boîte est formée par les 1er quartile q_1 et 3ème quartile q_3
- les moustaches sont définies par les valeurs observées les plus extrêmes dans l'intervalle $[q_1 - 1.5(q_3 - q_1), q_3 + 1.5(q_3 - q_1)]$
- $\circ$: valeurs extrêmes hors des moustaches

Représentation graphique d'une variable quantitative

Histogramme : surface de chaque barre proportionnelle à l'effectif de la classe d'observations correspondante



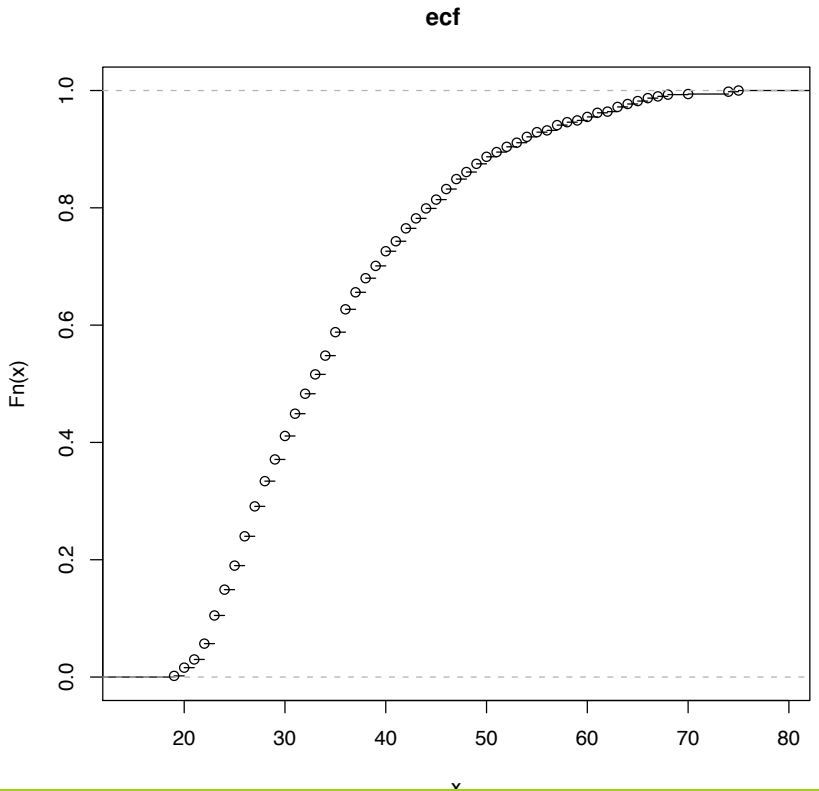
Rq : si la variable est quantitative discrète, les "barres" sont remplacées par des "bâtons".

Représentation graphique d'une variable quantitative

Fonction de répartition empirique :

$$F_n(x) = \frac{N_x}{n}$$

où $N_x = \#\{X_i : X_i \leq x, 1 \leq i \leq n\}$ est le nombre de données inférieures ou égales à X.



Plan

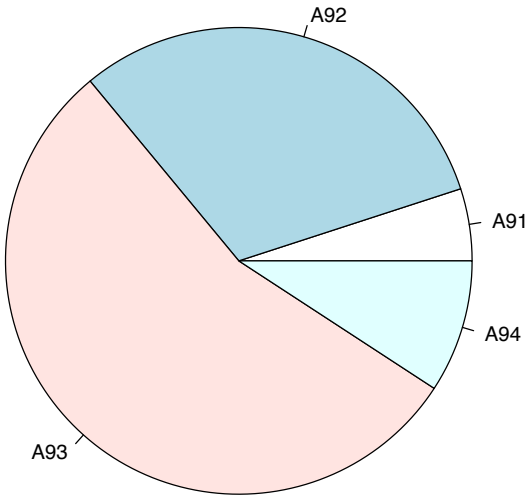
Résumé numérique d'une variable qualitative

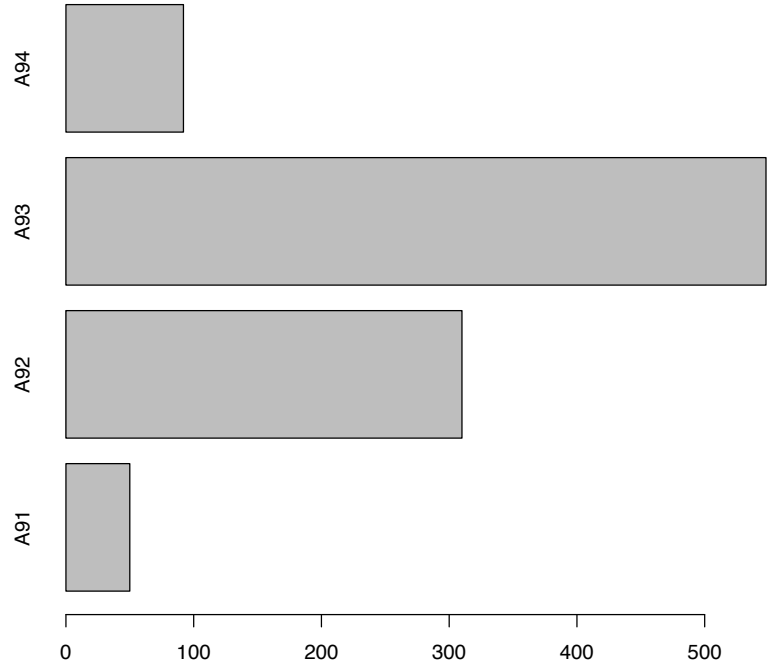
Si X est une variable aléatoire qualitative prenant les modalités $\{m_1, \ldots, m_p\}$ on s'intéresse à

- $N_j = \#\{X_i : X_i = m_j, 1 \leq i \leq n\}$: le nombre d'occurrences (effectif) de la modalité m_j dans l'échantillon
- $F_i = \frac{N_i}{n}$: la **fréquence** de la modalité m_j dans l'échantillon

Représentation graphique d'une variable qualitative

Diagramme en camembert (*pie-chart*, gauche) et en barres (droite)





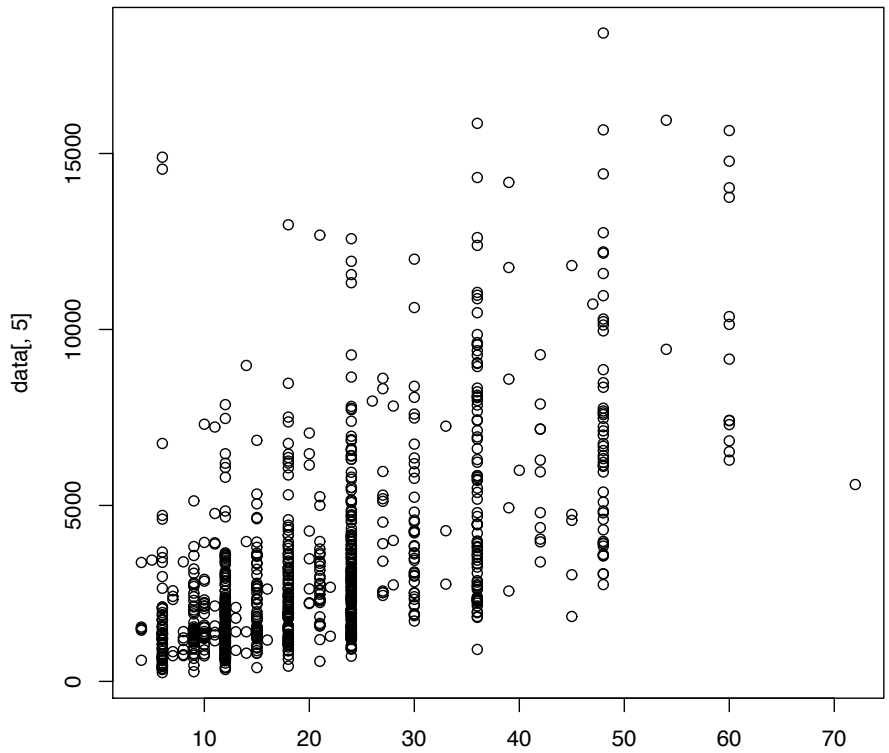
Plan

Décrire la liaison entre deux variables quantitatives

Numérique : le coefficient de corrélation linéaire

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}.$$

Graphique : le nuage de points :



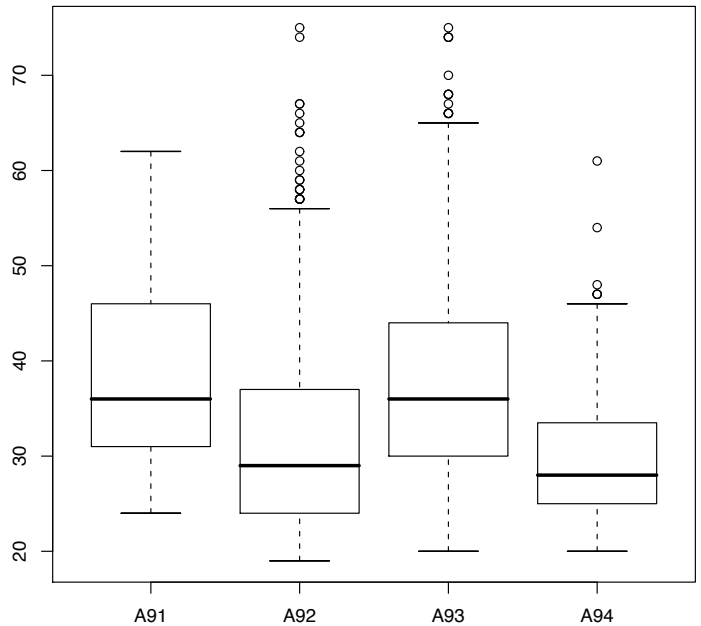
Décrire la liaison entre Y quantitative et X qualitative

Numérique : le rapport de corrélation :

$$R_{Y|X} = \sqrt{V_X^2 / V^2}$$

où $V_X^2 = \frac{1}{n} \sum_{j=1}^{R} N_j (\bar{Y}_j - \bar{Y})^2$ exprime la part de variabilité due à X
($\bar{Y}_j$ étant la moyenne des Y observés lorsque $X = j$ème modalité)

Graphique : le boxplot :



Décrire la liaison entre deux variables qualitatives

X : modalités $m_1, \ldots, m_R$
Y : modalités $o_1, \ldots, o_C$

	o_1	$\cdots$	o_c	$\cdots$	o_C	sommes
m_1	N_{11}	$\cdots$	N_{1c}	$\cdots$	N_{1C}	$N_{1.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
m_r	N_{r1}	$\cdots$	N_{rc}	$\cdots$	N_{rC}	$N_{r.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
m_R	N_{R1}	$\cdots$	N_{Rc}	$\cdots$	N_{RC}	$N_{R.}$
sommes	$N_{.1}$	$\cdots$	$N_{.c}$	$\cdots$	$N_{.C}$	n

Table : Table de contingence

Décrire la liaison entre deux variables qualitatives

Mesure de liaison

- le χ^2 (non borné):

$$\chi^2 = \sum_{r=1}^{R}\sum_{c=1}^{C} \frac{\left(N_{rc} - \frac{N_{r.}N_{.c}}{n}\right)^2}{\frac{N_{r.}N_{.c}}{n}} = n\left[\sum_{r=1}^{R}\sum_{c=1}^{C} \frac{N_{rc}^2}{N_{r.}N_{.c}} - 1\right]$$

- le $\Phi^2 = \frac{\chi^2}{n}$ (dépend encore de C et de R)
- le C de Cramer ($\in [0,1]$) $C = \sqrt{\frac{\Phi^2}{inf(R,C)-1}}$
- le T de Tschuprow ($\in [0,1]$) $T = \sqrt{\frac{\Phi^2}{(R-1)(C-1)}}$

Plan

Notion d'estimation

X une variable aléatoire
θ un paramètre caractérisant X dans une population donnée.

Ex :

- θ : la moyenne de X, $E[X] = \mu$
- θ : la variance de X, $V(X) = \sigma^2$
- θ : la proportion de telle modalité, lorsque X est qualitative...

Notion d'estimation

X une variable aléatoire
θ un paramètre caractérisant X dans une population donnée.

Ex :

- θ : la moyenne de X, $E[X] = \mu$
- θ : la variance de X, $V(X) = \sigma^2$
- θ : la proportion de telle modalité, lorsque X est qualitative...

Généralement, θ n'est pas connu car on ne dispose pas de mesure de X sur toute la population
$\Rightarrow$ on va chercher à estimer θ à partir d'un échantillon $X_1, \ldots, X_n$

$\Rightarrow$ estimation ponctuelle : donner une valeur approchée de θ

$\Rightarrow$ estimation par intervalle : donner un intervalle encadrant la vraie valeur de θ avec une certaine probabilité

Notion d'estimateur

Estimateur

Une statistique T fonction de l'échantillon : $T(X_1, \ldots, X_n)$
dont on appréciera les qualités suivantes :

- convergent : $T(X_1, \ldots, X_n) \overset{n \to \infty}{\longrightarrow} \theta$
- sans biais : $E[T - \theta] = 0$
- de variance $V(T) = E[(T - \theta)^2]$ minimale

Plan

Estimateur de l'espérance et de la variance

Estimateur de l'espérance

La statistique $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ est le meilleur estimateur de μ

- convergent : $\bar{X} \overset{n \to \infty}{\longrightarrow} \mu$
- sans biais : $E[\bar{X}] = \mu$
- de variance $V(\bar{X}) = \frac{\sigma^2}{n}$ minimale

Estimateur de l'espérance et de la variance

Estimateur de l'espérance

La statistique $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ est le meilleur estimateur de μ

- convergent : $\bar{X} \overset{n\to\infty}{\longrightarrow} \mu$
- sans biais : $E[\bar{X}] = \mu$
- de variance $V(\bar{X}) = \frac{\sigma^2}{n}$ minimale

Estimateur de la variance

La statistique $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ est le meilleur estimateur de σ^2

- convergent : $S^2 \overset{n\to\infty}{\longrightarrow} \sigma^2$
- sans biais : $E[S^2] = \sigma^2$
- de variance minimale

Rq : V^2 vu précédemment est biaisé $E[V^2] = \frac{n-1}{n}\sigma^2$

Estimateur d'une proportion

On s'intéresse à la proportion p d'individus ayant une certaine caractéristique dans une population.

Estimateur d'une proportion

$$F = \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n},$$

où X_i est une v.a. de Bernoulli de paramètre p, définie par :

$$X_i = \begin{cases} 1 \text{ si l'individu i possède la caractère C} \\ 0 \text{ sinon.} \end{cases}$$

Plan

Intervalle de confiance sur l'espérance

Loi de $\bar{X}$

- $X_i \sim \mathcal{N}(\mu, \sigma^2) \Rightarrow \bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$
- lorsque les X_i ne sont pas gaussiens
  Théorème centrale limite : $\bar{X} \overset{n\to\infty}{\longrightarrow} \mathcal{N}(\mu, \frac{\sigma^2}{n})$

Intervalle de confiance sur μ

$IC_{1-\alpha}(\mu)$

loi normale ou $n \geq 30$

σ^2 connue	σ^2 inconnue
$[\bar{x} + u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} - u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}]$	$[\bar{x} + t_{n-1,\frac{\alpha}{2}} \frac{S}{\sqrt{n}}, \bar{x} - t_{n-1,\frac{\alpha}{2}} \frac{S}{\sqrt{n}}]$

Intervalle de confiance sur la variance

Loi de $\bar{X}$

- $X_i \sim \mathcal{N}(\mu, \sigma^2) \Rightarrow \frac{n-1}{\sigma^2} S^2 \sim \chi^2_{n-1}$

Intervalle de confiance sur σ^2

$IC_{1-\alpha}(\sigma^2)$

loi normale

μ connue

$$\left[\frac{nV_\mu^2}{\chi^2_{n,1-\frac{\alpha}{2}}}, \frac{nV_\mu^2}{\chi^{2n,\frac{\alpha}{2}}}\right]$$

μ inconnue

$$\left[\frac{(n-1)s^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}, \frac{(n-1)s^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right]$$

Intervalle de confiance sur une proportion

Loi de F

- $F = \frac{\sum_{i=1}^{n} X_i}{n}$ avec $X_i \sim \mathcal{B}(p) \Rightarrow nF = \sum_{i=1}^{n} X_i \sim \mathcal{B}(n, p)$
- si n est suffisamment grand ($np > 5$ et $n(1-p) > 5$) :

$$F \sim \mathcal{N}(p, \frac{p(1-p)}{n})$$

Intervalle de confiance sur p

$$IC_{1-\alpha}(p)$$

$$np > 5 \text{ et } n(1-p) > 5$$

$$[f + u_{\frac{\alpha}{2}} \sqrt{\frac{f(1-f)}{n}}, f - u_{\frac{\alpha}{2}} \sqrt{\frac{f(1-f)}{n}}]$$